П.Ю. Довнар (Минск, ИП «Инвеншион Машин»)

АВТОМАТИЧЕСКИЙ АНАЛИЗ ТЕКСТА НА КИТАЙСКОМ ЯЗЫКЕ

Автоматическая обработка текста — это преобразование текста на искусственном или естественном языке с помощью ЭВМ.

Данное исследование посвящено разработке лингвистического процессора китайского языка, внедренного в состав промышленного многоязычного лингвистического процессора системы автоматизации инженерии и управления знаниями Goldfire Innovator компании «Инвеншион Машин». Его функциональность прежде всего направлена на реализацию автоматического лингвистического анализа на всех уровнях глубины естественных языков.

По распространенности китайский язык занимает первое место в мире. Ввиду того что на этом языке создается огромное количество технической литературы (патенты, инструкции и т.д.), растет потребность в налаживании межъязыковой коммуникации, что делает актуальным создание программ по автоматической обработке китайского языка.

Китайский язык пользуется иероглифической письменностью. Каждый иероглиф обозначает отдельный слог и отдельную морфему. Доминирует двусложная (двуморфемная) норма слова. В связи с развитием терминологии растет число более чем двусложных слов. Словообразование осуществляется за счет словосложения, аффиксации и конверсии. Формообразование представлено главным образом глагольными видовыми суффиксами. Аффиксы немногочисленны, в ряде случаев факультативны, имеют агглютинативный характер. Агглютинация не служит выражению отношений между словами, и строй китайского языка остается изолирующим. Синтаксис характеризуется номинативным строем, грамматически значимым порядком слов, определение всегда в препозиции. Китайский язык имеет развитую систему сложных предложений, образуемых союзным и бессоюзным сочинением и подчинением [2].

Автоматическая обработка китайского языка по структуре не отличается от обработки английского, немецкого, французского языков: входной текст проходит последовательные этапы формализации и структуризации. Можно выделить следующие основные уровни анализа текста: графематический, лексико-грамматический, синтаксический и семантический [3]. Большинство проблем, с которыми нам пришлось столкнуться, связано с первыми тремя из перечисленных уровней. На начальных уровнях анализа текста нам приходилось искать решения, адаптированные под специфику и особенности именно китайского языка.

На уровне **графематического анализа** текста на китайском языке основная проблема связана с отсутствием в тексте пробелов, в связи с чем трудно определить границы слов. Связь между частями сложного слова в китайском языке осуществляется путем свободного примыкания, не получая формального выражения посредством специальных морфологических показателей. При этом большое количество слов, образованных таким способом, может быть интерпретировано как цельные понятия и как синтаксические конструкции.

Исследования показали, что при решении данной проблемы с точки зрения промышленного характера ЛП целесообразно руководствоваться «прагматическим» принципом: разбивать слова так, чтобы они оптимально подходили для задач последующего анализа и обработки. Деление слов на слишком мелкие части приводит к увеличению многозначности и, соответственно, трудоемкости обработки. Игнорирование же смысловых частей слова, будь то полнозначные слова или отдельные морфемы, на практике приводит к снижению показателя полноты извлекаемой информации.

Таблица 1

Описание проблемы	Примеры	Решение
Многие глаголы представляют собой сочетание акции и объекта	**走路**–идти: 走–идти, 路–путь. **说话**–говорить: 说–говорить, 话 — речь.	Если глагол обозначает единое действие и если у объекта нет зависимых слов, то такие слова объединяются в одно
Существует несколько так называемых «морфем», состоящих из одного иероглифа. Эти «морфемы» используются вместо полнозначных слов и образуют новые слова (эти морфемы можно считать своеобразными суффиксами).	**率** — «коэффициент»: 生产**率** — производительность (коэффициент производительности), **机** прибор: 发射**机** — излучатель, 录音**机** — проигрыватель, 收音**机** — радиоприёмник	Составлен список таких морфем с указанием их лексических значений; образованные с их помощью слова распознаются как одно; при необходимости объединенные слова могут быть разбиты на этапе семантического анализа, и из них может быть извлечена необходимая информация
Термин состоит из нескольких полнозначных слов	**精神分裂症** — — шизофрения: 精神 — сознание, 分裂 — распад, 症 — болезнь.	При работе с терминами выделяются все значимые части.

В таблице 1 приводятся примеры и описываются некоторые проблемы, с которыми мы столкнулись на этапе графематического анализа текста. Представлены также их возможные решения.

Основная проблема на уровне **лексико-грамматического анализа** текста связана с классификацией слов. Для английского, немецкого, французского языков за основу такой классификации принято использовать разделение слов по частям речи [1]. Ряд исследователей отрицают наличие в китайском языке частей речи, аргументируя это отсутствием словоизменения в европейском понимании. Однако при выделении синтаксических и семантических отношений между словами невозможно обойтись без какой-либо предварительной лексико-грамматической классификации. Поэтому было принято решение о выделении традиционных частей речи (существительное, местоимение, прилагательное, глагол, наречие, числительное). За основу классификации была взята классификация частей речи в лингвистическом процессоре английского языка, разработанного компанией ранее. Разработанный лексико-грамматический классификатор включает в себя всего около 50 различных классов слов. Следует отметить, что данный классификатор включает в себя только те классы слов, точное автоматическое распознавание которых достижимо с опорой на локальный контекст слова. В случаях, когда для распознавания того или иного класса необходим анализ более широкого контекста, соответствующее решение должно приниматься ЛП на этапе синтаксического анализа.

В качестве тренировочной и тестировочной базы для реализации графематического и лексико-грамматического анализа текста используется разработанный нами эталонный корпус текстов, аннотированный лексико-грамматическими классами. Данный корпус содержит тексты различных жан-

ров и предметных областей в сбалансированной пропорции. Его объем составляет около 50 тыс. предложений (более 1.3 миллиона словоупотреблений).

Анализ указанного выше корпуса подтвердил, что в китайском языке показатель синтаксической транспозиции слов достаточно высок. Большое количество слов может использоваться в функции нескольких частей речи. Достаточно проблемными в этом плане и с точки зрения автоматической обработки текста являются следующие пары классов слов: существительное — глагол, прилагательное — существительное, прилагательное — наречие и предлог — глагол. В некоторых случаях частеречная принадлежность слова в определенном контексте в силу его синтаксической позиции и функции является однозначной:

比埃尼亚斯研究小组详细**检查**了已知人类基因的基因活性。*Исследовательская группа* ***проверила*** *жизнеспособность генов.*

要在这里进行严格**检查**。 *Здесь необходимо провести строгую* ***проверку****.*

С другой стороны, встречается много случаев, когда допустимы различные трактования и синтаксической роли слова, и, соответственно, его частеречной принадлежности:

用扫描仪进行监测。***Используя*** *сканер, осуществить контроль.* ***С помощью*** *сканера осуществить контроль).*

经分选清理取得海绵铁 。 ***Проведя*** *очистку, получить железо. Получить железо* ***посредством*** *очистки.*

Для того чтобы частично минимизировать указанную многозначность и, соответственно, повысить точность анализа текста было принято решение максимально избавиться от омонимии «предлог — глагол» путем однозначной классификации слов в тот или другой класс.

Графематический и лексико-грамматический этапы анализа были объединены в рамках единой стадии обработки текста. Такой подход позволяет быстро и эффективно находить среди большого количества всех возможных вариантов разбиения предложения на слова именно те, которые максимально вероятны с точки зрения лексико-грамматического анализа текста. В качестве статистической модели была использована модель Маркова. В дополнение к этому был разработан механизм распознавания незнакомых слов, а также механизм корректировочных правил. При таком подходе точность анализа на сегодняшний день составляет около 93%.

В качестве тренировочного и тестировочного материала используется разработанный нами эталонный корпус синтаксических деревьев. Его объем составляет около 4 тыс. предложений.

Проведенное исследование подтвердило, что технология автоматической обработки китайского языка не имеет принципиальных отличий по сравнению с технологией обработки английского, немецкого, французского языков, а полученный лингвистический процессор китайского языка может быть эффективно внедрен в состав системы автоматизации инженерии и управления знаниями Goldfire Innovator.